Forecasting Student GPA using Data from Kaggle

Khomotjo Mathabatha
School of Computer Science and Applied Mathematics
The University of the Witwatersrand
Johannesburg, South Africa
1432040@students.wits.ac.za

Dr. Ashwini Jadhav
Science Teaching and Learning Center
Faculty of Science
The University of the Witwatersrand
Johannesburg, South Africa
Ashwini.Jadhav@wits.ac.za

Dr. Ritesh Ajoodha
School of Computer Science and Applied Mathematics
The University of the Witwatersrand
Johannesburg, South Africa
ritesh.ajoodha@wits.ac.za

***Abstract*—Institutes need ways to improve the quality of the education system to achieve better results and reduce the failure rate. Education is very important in society and to the development of each country to improve the livelihood of the citizens. In this study, machine learning algorithms will be used to forecast students GPA and the highest accuracy was achieved using Naive Bayes which showed that the algorithm was a good model to use on our data. Different techniques were chosen and tested on the data found on Kaggle and evaluation metrics like F1 score, precision and recall were used to compare the performance of the techniques.**

I. Introduction

GPA is a very important measure for student's academic performance and is measured by taking the average of all assessments that the students were given throughout the year like assignments, projects, labs, tests and exams. Being able to find ways to improve the learning and teaching methods to improve student performance is very important to universities and machine learning algorithms are the most popular ways to evaluate and find useful information from their educational data. Machine learning algorithms will be extensively applied to the educational data to forecast student GPA to help identify students that might perform badly so that the teaching and learning approaches can be improved. Features that affect the GPA will be analysed and used to create a model that can forecast student GPA. [8].

Academic performance for students is difficult to assess since it is influenced by a range of factors including demographics, educational experience and other external influences. Machine learning approaches have recently become popular for investigating educational data and identifying hidden relevant patterns for forecasting student's grades. Previous studies predicted student's GPA using some of the techniques I will be implementing as they achieved a high accuracy and I will be using some of those techniques to see which will provide a high accuracy and also try to find the reasons why some of the techniques might perform poorly. The study will follow with the problem statement and the Literature review where past papers will be discussed including the data used and the results found in those studies. After the focus will on the methodology where the different methods used in the study will be discussed and explained. Then we will discuss our findings in the conclusion and also discuss some of the different ways our study and findings can be improved our in the future work section. Then we conclude with the reference of the studies that our paper is based on.

Author	Data	Features	Models
Aysha Ashraf	University data	Grades and activities	KNN, Naive bayes and SVM
Mehel Shah	UCI machine Learning Repository	Previous Marks and activities	Decision Tree ,SVM and Random Forest
Ihsan A Amra	Secondary School data	Previous marks	KNN and Naive Bayes
Vairachilai S	Kaggle	Marks ,attendance and past performance	SVM ,Naive Bayes and Decision Tree
Maria Koutina	University data	Behaviour	SVM, Random Forest , KNN and Naive Bayes

TABLE I
TABLE SHOWING SOME OF THE PREVIOUS STUDIES REFERENCED IN THIS PAPER.

II. Problem Statement

The failure rate at most institutes is high and finding ways to reduce it is very crucial to the education system of every country. This research can help by identifying students who are likely to do well and those who might need immediate attention which can result in the reduction of the failure rate.

III. Literature Review

Table 1 shows some of authors of the papers used in this study with the methods they used in their papers. Some of the important features were also added with the source of the data set that they used and their result will be discussed.

Education is very important for creating the best lives for ourselves and those around us. Educational data is rapidly

increasing as the admissions in the universities increase [1], [6] and most of it remains unused [8]. With the growth in computer and information technology, a process was developed that enables us to use the huge educational data sets. Data mining is the process used to extract useful, hidden and relevant information from large data sets. Machine learning is the most popular method used [11] in creating models to enhance and explore academic performance and there are a lot of studies investigating the application of Machine Learning algorithms to the educational data sets to try and reduce dropouts, failure rate and improve teaching by finding different and better methods of teaching.

There are two categories that all machine learning algorithms used for performance prediction fall into and the first category is classification algorithms which are the commonly used data mining technique that develop a model using pre-classified examples to classify the data into classes or categories. The other category is clustering which finds group of objects and classify them using their attributes so that objects with similar characteristics can be grouped together so that they can be easily differentiated with other groups by either behaviour or activities [3].

There a lot of different ways to acquire the needed data to test the models on but the most commonly used data is the university data [1],[2],[4],[7],[10] and the studies focused on the student academic performance part of the data set. Financial and personal data were never used in any of the studies. High school data [6] and the UCI Machine Learning Repository are other useful sources of data with Kaggle also being a useful source if access to educational data cannot be found as most institutes protect their student's data and sometimes it can be difficult to find the required data to test the algorithms. Apps [3] can also be a good source to get the student's databases through the Knowledge Discovery Databases.

Most of the data sets are huge and they have to be re-sampled by focusing on a certain years and courses to be able to test the models before applying it on the entire data set. Pre-processing the data is a very important part in the process of applying the techniques. The Waikato Environment for Knowledge Analysis workbench which is a java written data mining tool created in New Zealand, Waikato University and is mostly used [1]. The large data set can not be used directly as it has many attributes linked by foreign keys, primary keys and relations that hinder predictive algorithms. Some data has imbalance caused by outliers in the data from students who are under performing or over performing.

Data imbalance is a huge problem and many studies have been trying to tackle the problem [2]. The methods to use in reducing the imbalance depend on the data and is done by random over sampling of the majority class and over sampling of the minority class. Feature selection can also be used as it selects a subset of features that allow classifiers to reach their optimal performance [2]. The aim is to achieve the best data quality by using data reduction and selecting the best attributes without losing the quality of the data to reduce the memory requirement and computation complexity. The CRISP framework can also be used for mining academic data.

The huge amount of data used has a lot of features and knowing the ones that have an effect on student GPA is very important in creating a good model to forecast student's GPA. Past results [2-8],[10],[11] either from high school or previous grades are the most used and the most important feature that shows the student's potential and what they can achieve.

Demographic information [2],[4-6],[8] like age, gender and occupation is the second most used feature and plays an important role in understanding the student. There are other factors like student behaviour that help in understanding the reasons behind sudden drops in student marks as a result of them needing extra help from teachers and tutors, involvement in extracurricular activities [2],[4],[5],[11] that get them involved and helps them focus, the student's location from the institute and the medium of teaching [10] used also have some effect on student GPA. J Dhilipan et al 2021 [9] highlighted the importance of a psychometric analysis. Accuracy, recall and precision [1], [11] are the most important measures used in finding which algorithms perform better.

The models were created using features that had an effect on the student's GPA and Machine Learning algorithms were applied to the models with the lowest accuracy produced being 70.48% and the highest being 100%. The most used Machine Learning algorithms are Support Vector Machine, Decision Tree, K-Nearest Neighbour, Naive Bayes, Neural Networks, Linear Regression and Random Forest. Other algorithms used are RIPPER, J48 and Backpropagation. The algorithms performed differently depending on the data and models used but mostly because of their characteristics.

Support Vector Machine is a classification algorithm primarily used to solve problems by finding a hyperplane in an N-dimensional space that group the data points [5]. Decision Tree algorithms are best with attributes like past marks, behaviour, demographic information and activities as they have two phases. The first phase is called the preparing phase and the second phase is the prune phase which reduce the data by breaking it down into smaller data sets which are also broken down into smaller subsets that help in the creation of a simple flow chart of leaf nodes and internal nodes that make up the simpler tree and the most used are ID3 and C4.5 [4].

K-Nearest Neighbour is one of the most basic and accurate algorithm for performance prediction in machine learning because is capable of understanding non-linear patterns in data but it's implementation is very complex as it makes underlying assumptions about the data distribution using

euclidean distance. The memory data points in the method need to be in memory at runtime which lead to it being slow and costing a lot of memory but it is easy to implement and works well with large data sets [6].

Naive Bayes is a classification algorithm that predicts student performance using probability theory and makes assumptions to simplify the problem [4] and estimates the likelihood of something occurring given a set of data as input, it performs better when the data dimensions are high and adding new raw data at runtime has no effect on the results which makes it efficient and it performs better on large data sets. Neural Networks are modeled after the human neural system.

Linear Regression uses a logic function to create the model and predicts the results by handling threshold values using precision and recall and it is also divided into three(Binomial, Multinomial and ordinal). Random Forest produce exceptional outputs in most of its applications because it is versatile, simple and has the ability to perform both regression and classification.

In [11] Support Vector Machine, Decision Tree and Naive Bayes algorithms were used to estimate student performance and Naive Bayes performed with an accuracy of 77%. In [5] student performance was analysed using Random Forest, Decision Tree, Linear Regression and Support Vector Machine with Random Forest having the highest accuracy of 91.7% but after a few manipulations and creating a new model using Gradient Boasting the accuracy increased to 93.8%. In [3] only Decision Tree algorithms were used to investigate how to improve student performance using an Adaptive Hypermedia Architecture system to mine student data with the help of teachers and C4.5 performed better in processing the academic data. In [6] an experimental study was done to help the ministry of education by estimating student performance to help lectures improve student learning by doing evaluations using K-Nearest Neighbour and Naive Bayes with K-Nearest Neighbour performing better with an accuracy of 83.65% and the model can be used to help lectures give better advice and classify student by performance and also show key acceptance criteria in university acceptance. All these papers prove that features and data selection play an important in the performance of the methods.

A lot of models were created and different algorithms were applied to them which resulted in different accuracy depending on the features and data used which proved that feature selection affects the performance of the models significantly [1] and improve accuracy [2]. Found that past performance has a big influence on student performance [10] and Machine Learning algorithms can make predictions on Educational Data [7]. Naive Bayes has a strong relationship with features that affect student performance [6] and to further improve the accuracy of the models we can use a radical basis function [9], a boasting algorithm that uses hyper parameters with fine tuning to improve performance [5], data manipulation and an ensemble method [8] can also be used. This will help in finding the best algorithm to predict student GPA.

IV. Methodology

A. Data

The dataset used is from Kaggle and had 649 instances with 33 attributes but 15 were removed. The attributes that were removed included free time, going out, parents job, alcohol consumption and romantic life.

B. Methods

There are a lot of machine learning algorithms to be used in forecasting student GPA but this study focused on K-Nearest Neighbour, Linear Regression, Naive Bayes, Random Forest, Support Vector Machine and Decision Tree as they produced better results and were easy to implement.

K-Nearest Neighbour is a simple, robust and easy to use supervised machine learning technique that is implemented to resolve either classification or regression issues. Classification problems give a discrete value as output while regression problems release a real number as output. Supervised learning techniques are algorithms that find the appropriate results when given new unlabeled data by implementing the model learned using labeled input data. K-Nearest Neighbour is effective for large training data, robust to noisy training data on top of being simple to implement but it is very difficult to determine the value of K and its computation cost are very high as it calculates the distance between all the training data set. It is also versatile as it works on many problems but it gets very slow as the number of predictors and variables increase.

$$\sqrt{\sum_{i=1}^{k}(x_i - y_i)^2} \tag{1}$$

The equation to calculate euclidean distance

Decision Tree is a supervised learning Artificial intelligence technique applied to either classification or regression problems but is it used in classification problems mostly. It is a tree structured classifier where the inside nodes are for the features of the data set, branches are for the rules on the decisions and each leaf represent an outcome. The Class and Regression Tree Algorithm (CART) is used to build the tree which basically asks questions and depending on the solutions, usually a Yes or a No the tree is split into sub-trees mimicking human thinking ability and leads to them being easily understood . They are easy to understand as they follow the same human thinking as people do in life, which helps us to consider all possible outcomes of a problem and requires less data cleaning compared to other algorithms, but it is complex and leads to over-fitting as it has many layers

and its computation complexity increases as the number of class labels increase. The leaf and branch are created using the folowing entropy equation :

$$\sum_{i=1}^{c} -P_i \log_2 P_i \tag{2}$$

If the entropy is zero then we have a leaf but if it is greater than zero we have a branch that needs further splitting.

Naive Bayes is one of the most used supervised learning techniques constructed on the idea of the Bayes theorem used to resolve classification issues utilizing a high extent of the training data set. It is one of the most effective and simple classification techniques which help in creating very fast Artificial intelligence models that make quick predictions by assuming that each occurrence of a certain feature has no relationship with the occurrence of another feature. It is fast and simple to implement in forecasting a class of data sets and it is used for either binary or multi-class categorization, but it has to learn all relations between its features as it assumes that they are unrelated.

Support Vector Machine is one of the most used supervised learning techniques and works in both classification and regression problems but mostly regression problems and it creates a best decision boundary that segregate dimensional space into classes so that all the new data points will be put into the correct class in future and picks the extreme vectors that assist in making the best decision boundary and it checks for the hyperplane that is able to differentiate between two classes and the formula for a hyperplane is

$$W^T x = 0 \tag{3}$$

Linear regression is an Artificial intelligence algorithm and it is one of the most picked algorithms.It is popular for it's use of the statistical method for predictive analysis, it makes prediction for continuous and numeric variables and makes any linear relation between a dependent variable and one or more independent variables visible. The formula is given by

$$Y_i = f(X_i, \beta) + e_i \tag{4}$$

where :
Y is the dependent variables

f is the function

β is the unknown

e is the error term

X is the independent variables

Random Forest is a popular Artificial intelligence process that belong to the supervised machine learning algorithms used for either classification problems or regression issues and it is constructed on the idea of an ensemble learning algorithm which is a procedure of putting together a lot of classifiers to solve a difficult issue to enhance the performance of the model. It has a lot of decision trees on different subsets of the retrieved data set and uses the mean to enhance the predictive accuracy of the entire data set. The algorithm has a better time complexity when compared with the other supervised learning algorithms, it produces the results at a high accuracy and maintains the accuracy even when a big part of the data is missing. It has the capability to handle huge data sets at high dimensionality and improves the predictive accuracy of the model which help avoid the possibility of over-fitting.

In K-Nearest Neighbour, the first step is to select the number of K neighbours and determine the euclidean distance mathematically for all the K neighbours. The second step is to extract the K nearest neighbours found using the length of a line segment between two points and the third step is to find the total number of data points in each class of the K neighbour. The final step is to allocate all the new data points to a class where the total number of neighbours is at a maximum and the model will be ready to be used. To use the model we load our training data and choose K by using the number of neighbours we have selected. Then for every instance of our data we determine the distance between our present instance and the query instance in our data set mathematically and combine the length and the index of the instance to any ordered array, collection or list. We then arrange the distances, starting with the smallest and ending with the largest in our ordered array, collection or list of distances and indices. After we pick the first K instances from the arranged array, collection or list and get the labels for all the K entries that were picked and output the mean of all the K labels.

Naive Bayes is where the data is turned into frequency tables and divided into two to create a feature matrix and a response vector. The response vector holds the output or prediction of each vector of the feature matrix while the feature matrix holds all the vectors of the data set where each vector consists of the values of the dependent features. Then we create the likelihood table by locating the probabilities of each given feature and calculate the posterior probability using the Bayes theorem.The Bayes theorem formula is

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

In Random Forest the algorithm functions in two phases, the first phase creates a random forest by putting together N decision trees and the second phase makes predictions for all the trees in the initial phase. It starts by selecting a random R points from the training data set and creates a decision tree using the selected data points. We then pick the number N as the number of the decision trees we want to make and replicate the process for all the training data. For all the new data points we make predictions for all the decision trees by assigning all the new data points to an instance that has the most votes as the winner.

We apply all the models to our data by first pre-processing our data and fitting each algorithm to the training set and let it predict the results as it learns for future use on big data sets like the educational data of an institute. Then we test the accuracy of the result using our evaluation metrics and visualize the test result for the data set. The algorithms were chosen because they performed better in previous studies and are good for our training data.

V. Results

Table 2 shows the results where the evaluation metrics such as accuracy, recall, precision and F1 score are used to measure the performance of the algorithms with Naive Bayes having the best accuracy while Decision tree and KNN had the worst accuracy. Naive bayes also has the best precision and F1 score with SVM having the best recall.

The formula to calculate the evaluation metrics :

$$F1score = 2 * \frac{Precision * Recall}{Precision + Recall}$$

$$Recall = \frac{TP}{TP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Accuracy = \frac{TP + TN}{TP + FN + FP + TN}$$

where :

TN is True Negative

TP is True Positive

FN is False Negative

FP is False Positive

Method	Precision	Recall	F1-score	Accuracy
Naive bayes	0.44	0.40	0.42	0.61
SVM	0.22	0.55	0.32	0.57
Linear Regression	0.15	0.44	0.23	0.50
KNN	0.25	0.17	0.2	0.40
Decision Tree	0.15	0.22	0.18	0.39

TABLE II
TABLE SHOWING THE RESULTS OF THE STUDY.

VI. Conclusion

The best algorithm to use in forecasting student GPA is Naive Bayes for the data set used in this study but Decision tree along with KNN are the worst. The results in the study are below what was found in other papers because the were more features used but machine learning algorithms are the best techniques to forecast student's GPA as all algorithms were able to predict the GPA of students using the training data.

VII. Future Work

In future I would like to test the algorithms on a huge data set like the University data and improve my results by sampling the data either under sampling or oversampling. To also improve the results I would like to use boasting algorithms and ensemble methods.

Ensemble methods also called Model combiners or Committee methods, are machine learning methods that use different algorithms that complement each other and use the power of those methods to improve it's accuracy in forecasting the results required. The algorithms are chosen if they are competent and the accuracy is higher than of any individual method.

A boasting algorithm improve the forecasting accuracy by training a family of weak models which compensate for each others weakness. The most known are Adaboast and Gradient algorithm and it forecasts by first applying equal weight/attention to every observation and in the second step it will focus the next learning algorithm to the forecasting errors of the previous learning algorithm and it will continue for the remaining algorithms while giving priority to the prediction errors of prior learning algorithms and will stop when a higher accuracy is achieved

Acknowledgment

I would like to thank Dr Ritesh Ajoodha and Dr Ashwini Jadhav for their help and support throughout the process of my research and being with me providing all the help I required and sharing their knowledge which made the process better.

References

[1] Havana Agraval and Harshil Mavani, "Student Performance Prediction Using Machine learning," Internationa Journal of Engineering Research and Technology, pp. 122–127, 2017.

[2] Abdulmonem A.A Ahmed , Aybaba Hancerliogullari and Yasemin Gultepe , " Classification Techniques to Predict Student's Performance of Higher Institute of Medical Sciences El-Shati" , IJESC , vol 10(6) , 2020.

[3] Maria Koutina and Katia Lida Kermanidis, " Predicting Postgraduate Student Performance Using Machine Learning Techniques ", International Federation for Information Processing , vol 364.9 , 2011 , pp. 159–168

[4] Aysha Ashraf , Sajid Anwer and Muhammad Gufran Khan," A Comparative Study of Predicting Student's Performance By Use of Data Mining Techniques " , American Scientific Research Journal for Engineering Technology and Science , vol44.1, 2018 ,pp 122–136

[5] Mehil B Shah and Maheeka Kaisha and Yogesh Gupta," Student Performance Assessment and Prediction System Using Machine Learning ", International Conference on Information System and Computer Networks, vol 4, 2019 , pp. 386–390

[6] Ihsan A.Abu Amra and Ashraf Y.A Maghari ," Students Performance Prediction Using KNN and Naive Bayesian " International Conference on Information Technology , vol 8,2017 ,pp. 909–913

[7] Boran Sekeroglu , Kamil Dimilier and Kubra Tuncal , "Student Performance Prediction and Classification Using Machine Learning Algorithms " ICEIT , 2019

[8] Muhammad Imran , Shahzad Latif , Danish Mehmood and Muhammad Saqlain Shah ," Student Academic Performance Prediction Using Supervised Learning Techniques " International Journal of Engineering and Techniques , vol 14.4, 2019

[9] J Dhilipan , N Vijayalaksi , S Suriya and Arockiya Christopher, " Prediction of Students Performance Using Machine Learning" , IOP Conference Series: Material Science and Engineering , vol 1055 , 2021

[10] K Govindasamy and T Velmurugan , " A Survey on the Result Based Analysis of Student Performance Using Data Mining techniques " International Journal of Data Mining Techniques and Applications , vol 4.1, 2015 pp 91–95,

[11] S Vairachilai , Avvari Sai Saketh and Gnanajeyaraman R," Students's Academic Performance Prediction Using Machine learning Aproach ", International Journal of Advanced Science and Technology, vol 29.9, 2020, pp 6731–6737